Defining the BTK-dependent chromatin-informed gene regulatory network in chronic lymphocytic leukemia

Zhiquan Wang[1], Weiguo Han[1], Qianqian Guo[2], Heather C. Darby[1], Sutapa Sinha[1], Chuanhe Yu[3], Sameer A. Parikh[1], Neil E. Kay[1,4]

1, Division of Hematology, Department of Medicine, Mayo Clinic, Rochester, MN, 55905, USA.
2, Division of Gastroenterology and Hepatology, Mayo Clinic, Rochester, MN 55905, USA.
3, The Hormel Institute, University of Minnesota, Austin, MN 55912, USA.
4, Department of Immunology, Mayo Clinic, Rochester, MN, 55905, USA.

Correspondence: Zhiquan Wang, wang.zhiquan@mayo.edu or Neil E. Kay, kay.neil@mayo.edu

- **Supplementary materials and methods**
- **Supplementary figs 1-6**
- **Supplementary references**

Materials and Methods

Bulk RNA-seq library preparation and sequencing

Total RNA was isolated using Direct-zol RNA Kit (Zymo Research) according to the manufacturer's instructions. RNA integrity was assessed by Bioanalyzer, and samples with RNA integrity number > 9 were used for library preparation. Library preparation and sequencing were performed using the NovaSeq 6000 platform, paired-end 150 bp by Novogene (Sacramento, CA).

Bulk RNA-seq preprocessing and differential expression analysis

Adapter trimming was performed with fastp.[1] Reads were aligned to GRCh38/hg38 using STAR,[2] and gene-level counts were generated with featureCounts (Subread).[3] Count matrices were imported into R and analyzed with DESeq2.[4] Genes with low counts were filtered before model fitting; in our standard workflow, genes were retained if counts were >5 in at least 3 samples. Wald statistics were used to estimate significance, and Benjamini-Hochberg correction was applied to control false discovery rate (FDR). Genes meeting FDR < 0.05 and |log2 fold change| > 1 were considered significant unless otherwise indicated. Principal component analysis and sample-level visualizations were generated using ggplot2.

ATAC-seq library preparation, alignment, and peak calling

ATAC-seq libraries were prepared from 50,000 cells per condition and sequenced as paired-end reads on NovaSeq 6000. Raw reads were processed through the project ATAC-seq pipeline, aligned to the human reference genome GRCh38/hg38 using Bowtie2 and filtered to generate analysis-ready BAM files.[5] Low-quality, duplicate, mitochondrial, and other excluded reads were removed according to the project filtering workflow, and only filtered BAM files were used for downstream quantification. Peaks were called for individual samples using MACS2,[6] and a consensus master peak set was generated across samples in the ATAC pipeline. Read counts over the consensus peak set were quantified from paired-end filtered BAM files using featureCounts v2.0.3 with the parameters -T 32 -p -F SAF. Differential accessibility analysis was then performed on the consensus peak count matrix using DESeq2 for the relevant contrast, and significantly opening or closing peaks were defined using FDR < 0.05 and |log2 fold change| > 1**.**

Histone-mark profiling and promoter/enhancer annotation

To define promoter- and enhancer-linked regulatory compartments, we profiled H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K27me3, and H3K36me3 using CUT&Tag. Sequencing reads were processed as above using fastp, bowtie2, samtools, and bedtools. Peaks were called using SEACR with stringent thresholds of 0.05. In our standard workflow, promoter regions were defined as the intersection of H3K4me3 master peaks and H3K27ac master peaks, and enhancer-like regions were defined as H3K27ac-positive regions lacking promoter-associated H3K4me3 signal. Interval operations were performed with bedtools.[7]

Integration of RNA, ATAC, and chromatin-state information

To construct a chromatin-informed regulatory atlas, differential RNA expression and differential chromatin accessibility were integrated with baseline promoter/enhancer annotation. Regulatory elements were first classified by baseline chromatin context and then linked to candidate target genes (nearest gene with 1Mb). Each linked element-gene pair was annotated by chromatin compartment, direction of accessibility change, and direction of RNA response after acute BTK degradation. This yielded a directional regulatory map connecting BTK-responsive chromatin elements to BTK-responsive genes in a chromatin-state-aware manner.

Definition of BTK-maintained and BTK-released regulatory programs

Linked element-gene relationships were partitioned into directional programs according to their response to acute BTK degradation. Elements and linked genes that decreased after BTK loss

were assigned to the BTK-maintained arm, whereas elements and linked genes that increased after BTK loss were assigned to the BTK-released arm. Aggregate program gene sets were defined as the nonredundant union of genes assigned to each directional arm after removal of duplicated gene symbols.

TF assignment and TF-program construction

Candidate TFs were assigned to BTK-responsive regulatory elements by integrating motif enrichment linked to target-gene behavior and directional program membership. Motif analysis was performed using HOMER.[8] TF-centered programs were constructed by mapping motif-supported TFs to responsive peaks and then propagating those assignments through linked target genes within the maintained or released network arm. TF-specific program gene sets were therefore defined as linked target genes associated with each TF in either the BTK-maintained or BTK-released atlas. Only TF programs with sufficient mapped genes in the relevant expression matrix were retained for downstream analyses.

TF hub prioritization

To identify the highest-confidence TF hubs within the chromatin-informed atlas, TF programs were prioritized using a composite framework that considered program size, strength of linked-gene support, directional coherence between chromatin accessibility and RNA response, and motif evidence. Ranking and summarization were performed in R. Representative high-confidence hubs in the BTK-maintained arm included NFATC1, JUN, FOSL1, FOSL2, NFKB1, and TCF12, whereas representative hubs in the BTK-released arm included ASCL1, TCF3, TCF12, PTF1A, and ASCL2. TCF12 was retained as distinct maintained and released TF programs because it mapped to separate directional target-gene sets in the atlas. These TF hubs were subsequently carried forward into single-cell analyses to test whether the bulk-inferred regulatory architecture was preserved across malignant cell states.

Single-cell RNA-seq validation in primary CLL B cells

Longitudinal single-cell RNA-seq data from the public CLL dataset GSE111014 were used to evaluate whether the BTK-maintained and BTK-released programs defined by the bulk chromatin-informed atlas could be resolved in malignant CLL cells *in vivo*. Serial peripheral blood samples from four patients collected at baseline and on treatment were analyzed in R using Seurat. Raw or processed expression matrices were imported into Seurat, and cells were filtered on the basis of standard quality-control metrics, including nFeature_RNA, nCount_RNA, and mitochondrial transcript fraction, using thresholds defined in the original analysis scripts.

Data were normalized in Seurat, followed by variable-feature selection, principal-component analysis, graph-based clustering, and UMAP embedding. Cluster marker genes were identified with FindAllMarkers using the method specified in the analysis scripts.

Broad cell-type annotation was performed using canonical marker expression together with cluster-level marker summaries. A broad B/CLL compartment was first defined from the full dataset, and a stricter malignant CLL subset was then generated as a sensitivity analysis using the criteria implemented in the original workflow. These analyses identified clusters 0, 4, 5, and 8 as the principal CLL states used for downstream state-resolved analyses.

Bulk-derived BTK-maintained, BTK-released, and TF-centered gene sets were transferred to the single-cell dataset and scored at the cell level using Seurat::AddModuleScore. Module scores were then summarized at the cell, cluster, sample, and patient levels using data.table and dplyr. In the broad CLL compartment, aggregate BTK-maintained and BTK-released scores were summarized by sample and ordered by collection time to visualize patient-level longitudinal trends.

To define baseline state structure, BTK-maintained and BTK-released scores were compared across clusters 0, 4, 5, and 8. Cluster-level distributions were visualized with violin plots, and pairwise comparisons were performed using Wilcoxon rank-sum tests with Benjamini-Hochberg correction. To quantify longitudinal remodeling, treatment-associated change for each program was calculated relative to each patient's baseline sample, and these delta values were summarized at the cluster and sample levels. Baseline and delta summaries were visualized using violin plots, trend plots, and heat maps.

To evaluate whether the TF architecture inferred from the bulk chromatin-informed atlas was preserved at single-cell resolution, we performed a state-resolved TF hub analysis across clusters 0, 4, 5, and 8. High-confidence TF programs from the maintained and released arms were selected on the basis of atlas assignment and retained for single-cell scoring if their member genes were sufficiently represented in the single-cell expression matrix. For each retained TF program, cell-level module scores were summarized to patient-by-cluster-by-timepoint means. Two complementary summaries were generated: baseline TF-program structure across clusters and treatment-associated delta values relative to patient-matched baseline. Representative maintained TF hubs included NFATC1, JUN, FOSL1, FOSL2, NFKB1, and TCF12, whereas representative released TF hubs included ASCL1, TCF3, TCF12, PTF1A, and ASCL2. Because TCF12 appeared in both directional arms of the atlas, maintained and released TCF12 programs were analyzed separately.

To characterize the biologic identity of the principal malignant states, differential-expression summaries for clusters 0, 4, 5, and 8 were subjected to Gene Ontology Biological Process enrichment analysis using the packages specified in the original workflow. Enrichment results were filtered at adjusted $P < 0.05$ and summarized across clusters to identify shared and state-enriched functional themes.

Quantification and statistical analysis

All analyses were performed in R unless otherwise specified. Differential bulk RNA and ATAC analyses used DESeq2 with Benjamini-Hochberg multiple-testing correction. No statistical method was used to predetermine sample size. Investigators were not blinded to group allocation during data analysis.

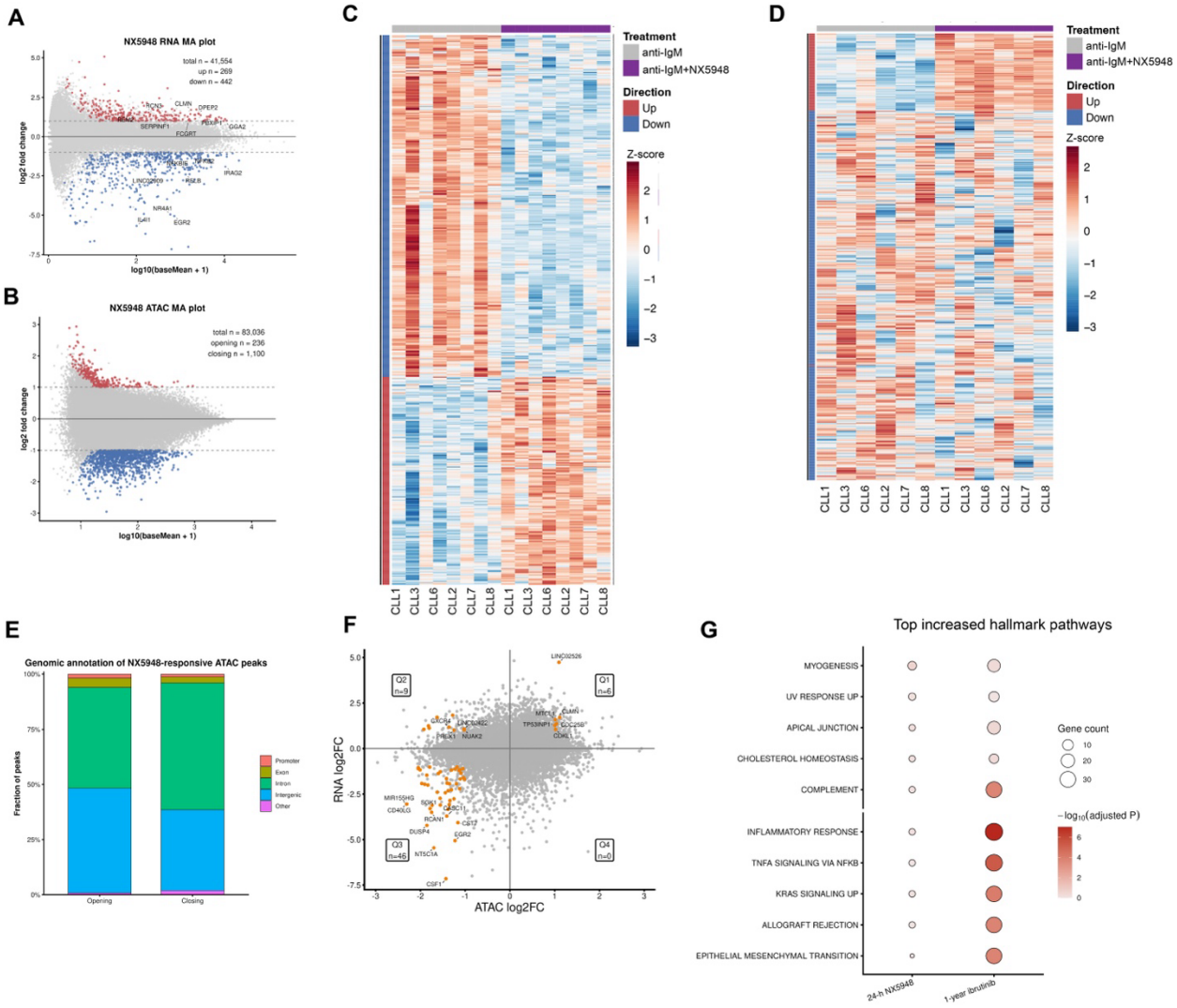

Figure S1. Expanded characterization of acute NX5948-induced transcriptional and chromatin responses in primary CLL B cells.
(A) MA plot of differential gene expression after acute NX5948 treatment. Significantly increased genes are shown in red, significantly decreased genes in blue, and non-significant genes in gray. Selected representative genes are labeled.
(B) MA plot of differential chromatin accessibility after acute NX5948 treatment. Opening peaks are shown in red, closing peaks in blue, and non-significant peaks in gray.
(C) Heatmap of significantly changed genes across primary CLL samples, displayed as row-scaled z-scores. Genes are grouped by direction of change, and samples are annotated by treatment condition.
(D) Heatmap of significantly changed ATAC peaks across primary CLL samples, displayed as row-scaled z-scores and annotated by treatment condition.
(E) Genomic annotation of NX5948-responsive ATAC peaks, shown separately for opening and closing peaks.
(F) Integrated RNA-seq and ATAC-seq quadrant plot comparing RNA log2 fold change and

ATAC log2 fold change for linked genes. Genes significant in both datasets are highlighted in orange.

(G) Signed increased Hallmark pathway comparison between acute NX5948 response and 1-year ibrutinib response.

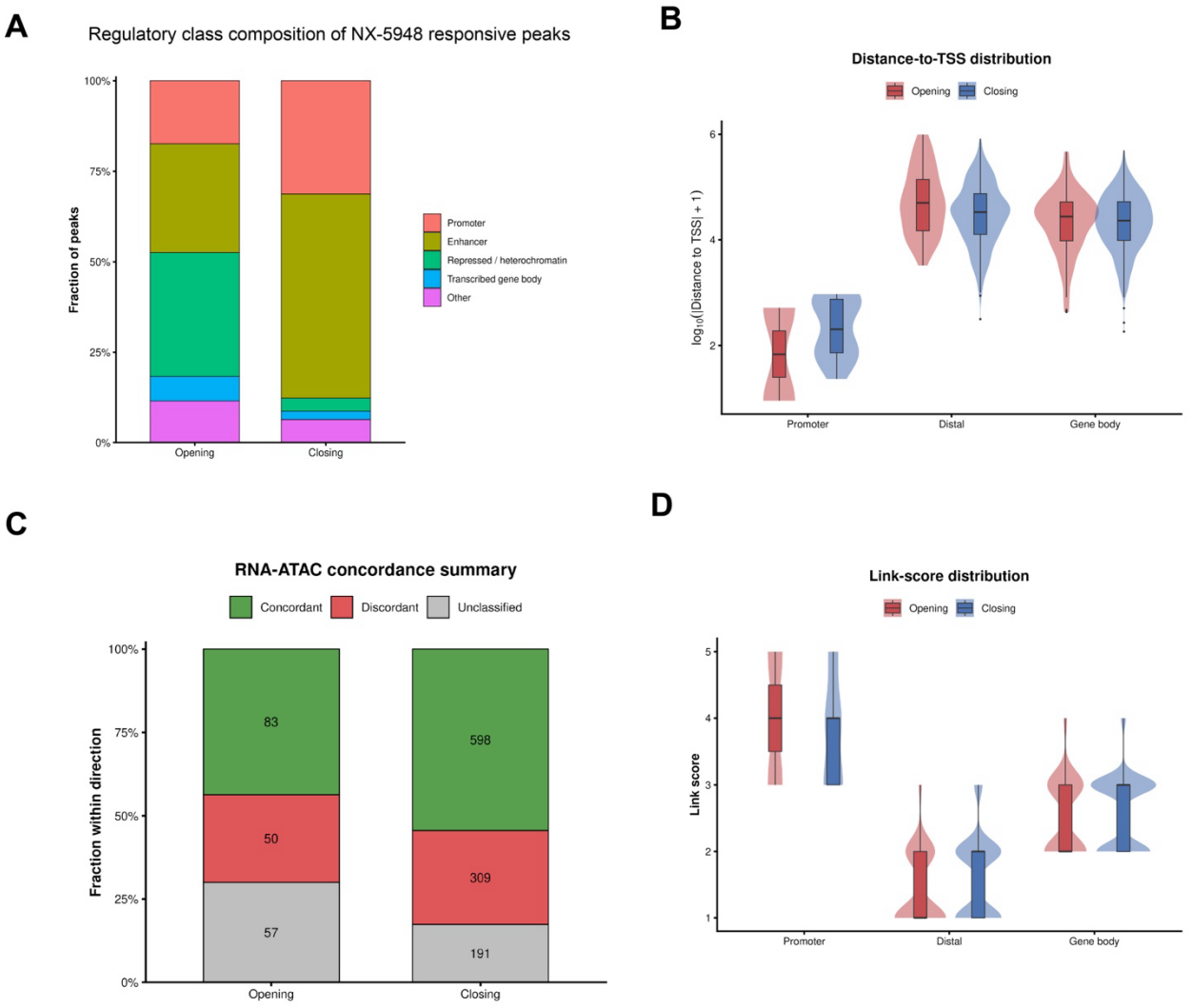

Figure S2. Supplementary characteristics of the NX5948 peak-to-gene linkage framework.
(A) Simplified regulatory-class composition of NX5948-responsive peaks. Chromatin states were collapsed into promoter, enhancer, repressed/heterochromatin, transcribed gene body, and other classes to highlight the major regulatory compartments underlying opening and closing peaks.
(B) Distance-to-TSS distributions of linked peak-gene rows stratified by link class and accessibility direction.
(C) Concordance of ATAC and RNA response among linked peak-gene rows, stratified by opening and closing peaks.
(D) The score distribution of linked peak-gene rows stratified by link class and accessibility direction.

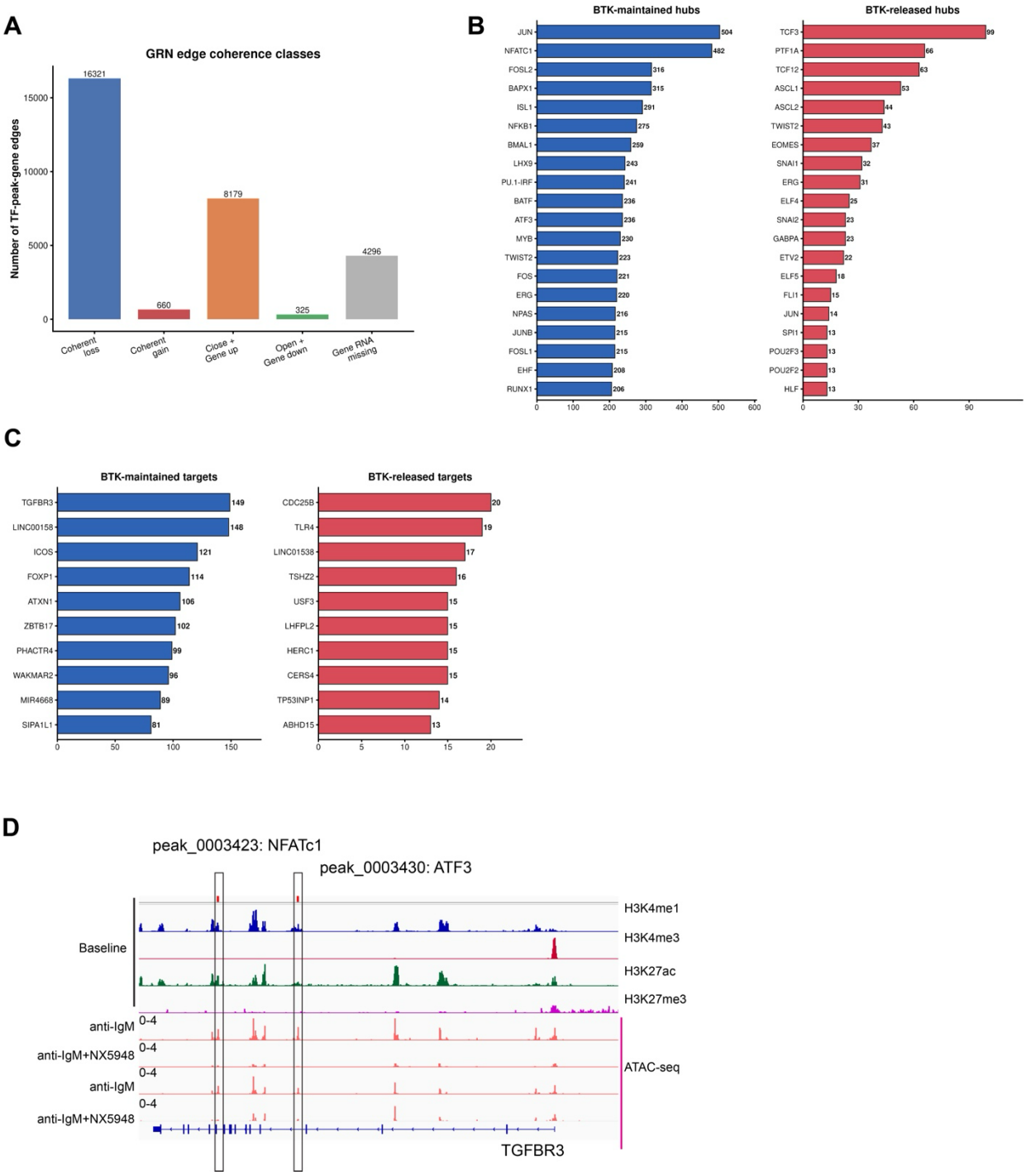

Fig. S3. Supporting structure of the chromatin-informed BTK regulatory atlas.
(A) Classification of candidate TF-peak-gene edges by coherence class, including coherent loss, coherent gain, close-plus-gene-up, open-plus-gene-down, and links with missing gene RNA information.
(B) Expanded ranking of top TF hubs in the BTK-maintained and BTK-released subnetworks, ordered by number of linked target genes after aggregating TF-assigned peak-to-gene edges by TF.
(C) Top target genes in the BTK-maintained and BTK-released subnetworks, ranked by number

of linked TF-peak inputs.
(D) Representative browser-track view of BTK-maintained regulatory elements at the **TGFBR3** locus. Highlighted **NFATC1**- and **ATF3**-assigned peaks are shown with baseline histone marks and ATAC-seq signal under anti-IgM and anti-IgM plus NX-5948 conditions, illustrating loss of accessibility at enhancer-like maintained elements after BTK degradation.

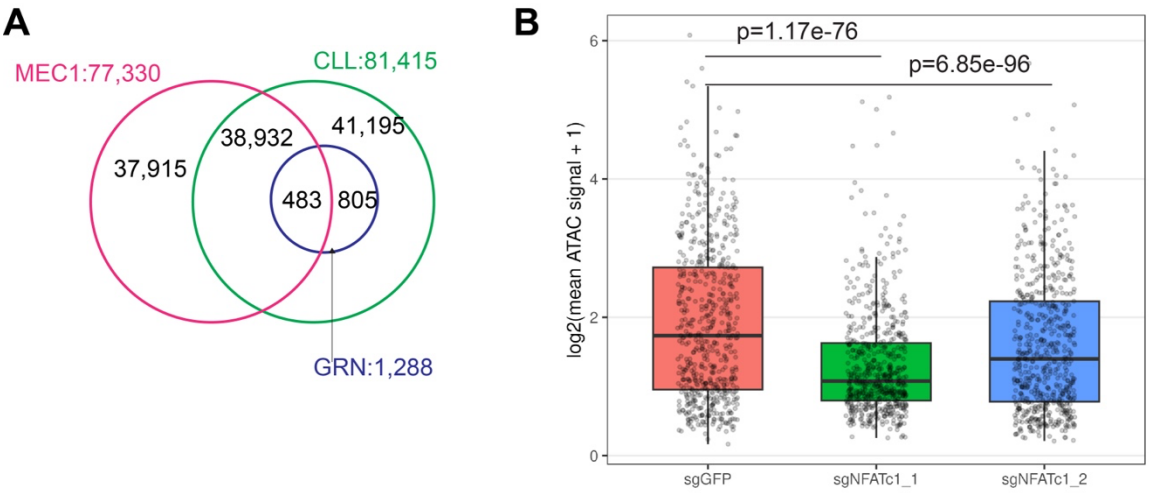

Fig. S4. MEC1 overlap with the CLL regulatory landscape and region-level quantification of NFATC1-dependent chromatin loss.
(A) Euler-style comparison of MEC1 ATAC peaks, GRN atlas peaks, and full CLL master peaks. Counts are based on combined-union genomic regions and show that the GRN atlas is fully nested within the full CLL master peak set, with a substantial subset represented in MEC1.
(B) Box plots with overlaid points showing log2(mean ATAC signal + 1) across the 598 atlas-defined BTK-maintained peaks for sgGFP, sgNFATc1_1, and sgNFATc1_2 MEC1 samples. P values indicate paired Wilcoxon comparisons versus sgGFP.

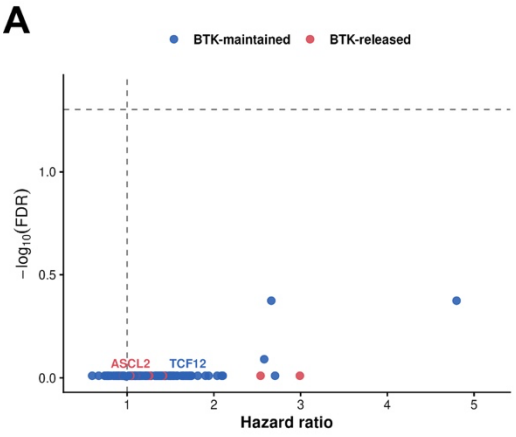

Fig. S5. BTK-responsive TF programs are not robustly associated with overall survival in the ICGC CLL cohort.

(A) Global TF-program OS scan across the chromatin-informed atlas. Each point represents one TF program, plotted by hazard ratio and −log10(FDR); selected representative programs are labeled. Blue, BTK-maintained; red, BTK-released. Dashed lines indicate HR = 1 and FDR = 0.05.

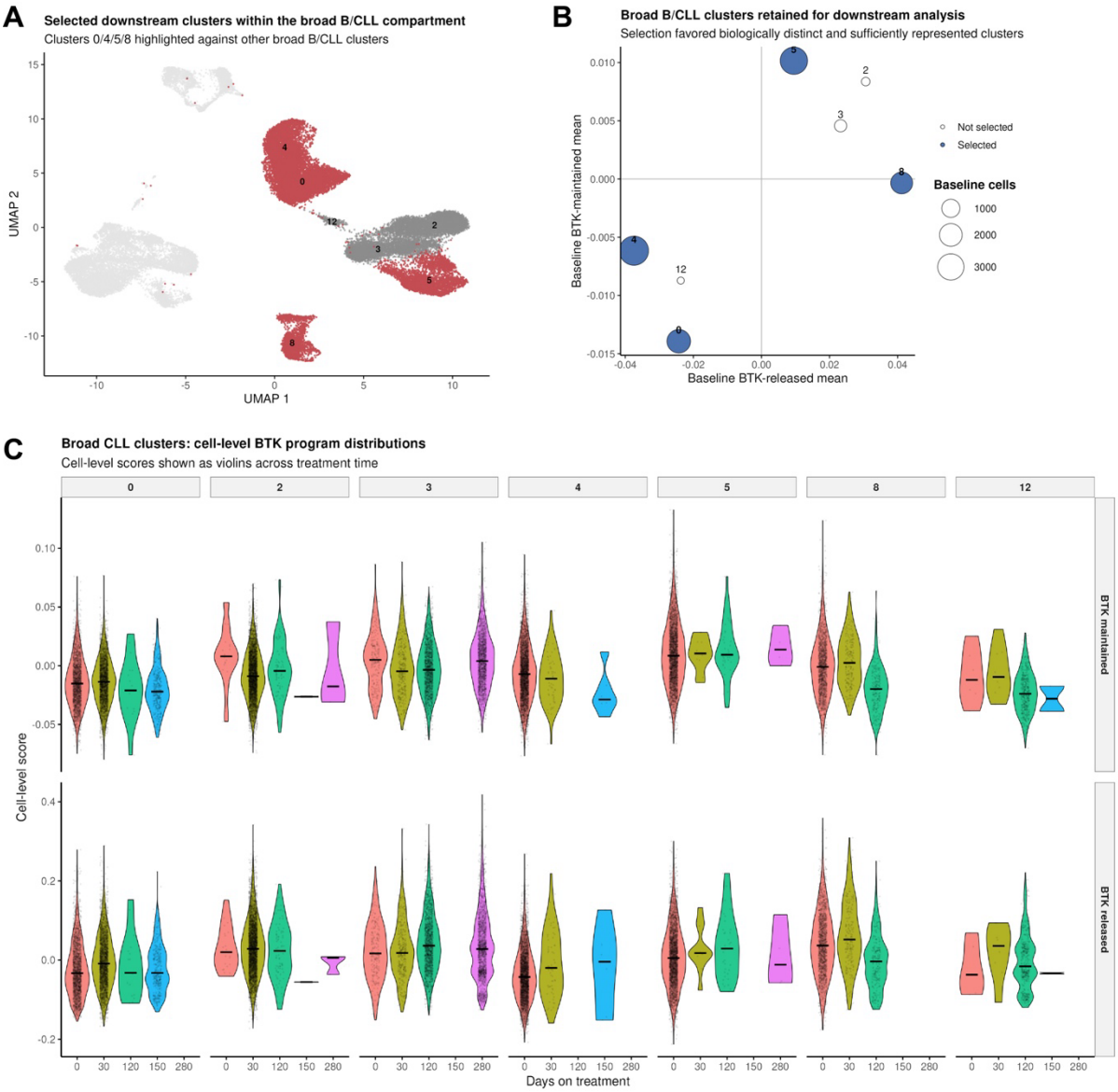

Fig. S6. Cluster selection and baseline single-cell architecture underlying the CLL validation analysis.
(A) UMAP highlighting the selected downstream CLL clusters within the broader B/CLL compartment.
(B) Cluster-selection framework used to nominate clusters 0, 4, 5, and 8 for downstream analysis.
(C) Baseline distributions of BTK-maintained and BTK-released program activity across the selected malignant states, showing that these programs are unevenly distributed across the compartment before treatment.

References

1. Chen S: **fastp 1.0: An ultra-fast all-round tool for FASTQ data quality control and preprocessing.** *Imeta* 2025, **4:**e70078.

2. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics* 2013, **29:**15-21.
3. Liao Y, Smyth GK, Shi W: **featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.** *Bioinformatics* 2013, **30:**923-930.
4. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome Biology* 2014, **15:**550.
5. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods* 2012, **9:**357-359.
6. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS: **Model-based analysis of ChIP-Seq (MACS).** *Genome Biol* 2008, **9:**R137.
7. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics* 2010, **26:**841-842.
8. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK: **Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities.** *Mol Cell* 2010, **38:**576-589.